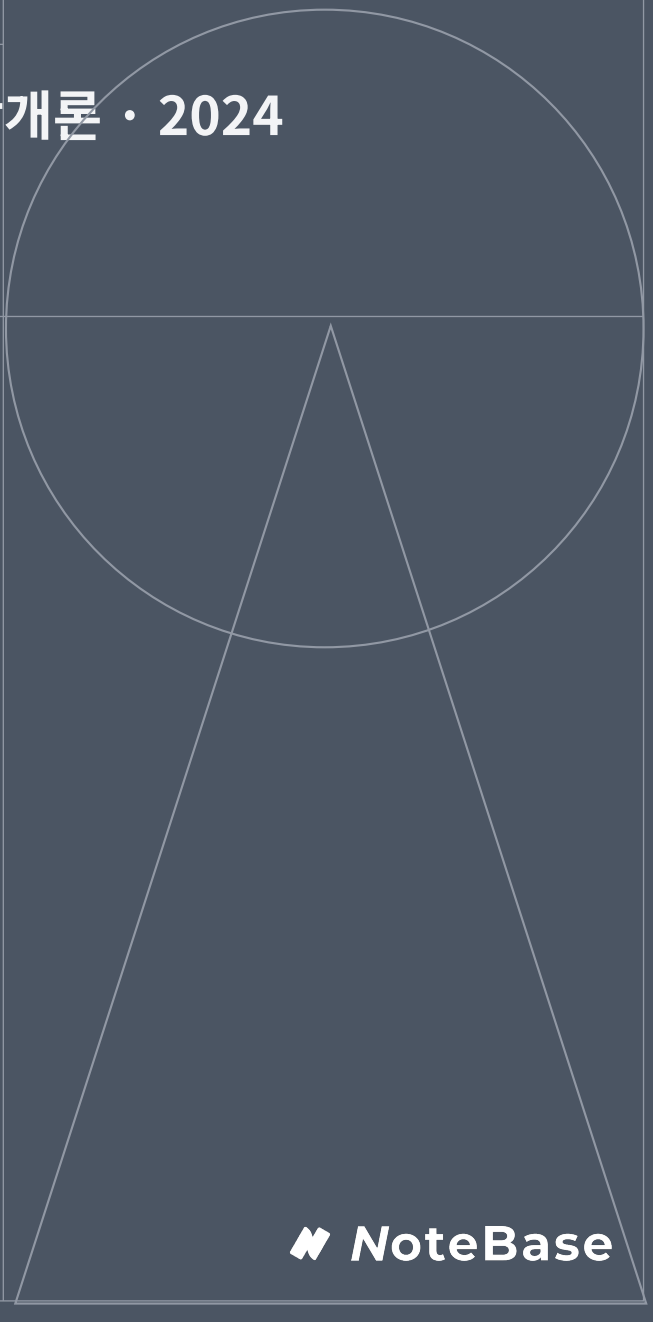
기출OX 해설집

공무원/국회직/9급 · 정보학개론 · 2024



NoteBase

기출OX 해설집

공무원/국회직/9급 · 정보학개론 · 2024

공무원/국회직/9급 · 정보학개론 · 2024

1. 다음 설명에 해당하는 학술지 영향력을 평가하는 지표는?

> ○ 한 학술지에 5년 동안 출판된 총 논문 수와 피인용횟수를 기반으로 측정한다.
> ○ 한 학술지가 다른 학술지에 비해 갖는 상대적 영향력을 평가하는 데 사용한다.
> ○ 페이지랭크 알고리즘(PageRank algorithm)과 유사한 방식으로 계산한다.

① SNIP(Source Normalized Impact per Paper)
② Eigenfactor
③ JIF(Journal Impact Factor)
④ CiteScore
⑤ Cited Half-Life

정답 2번

출제 포인트

이 문제는 학술지 영향력 지표를 계산 기간, 상대적 영향력, 인용 네트워크 반영 여부로 구별하는 문제이다.
제시된 세 특징을 모두 만족하는 지표는 Eigenfactor이다.

핵심 개념

Eigenfactor: 최근 5년간 학술지 인용 네트워크를 바탕으로 학술지의 총 영향력을 평가하며, 영향력 있는 학술지에서 받은 인용에 더 큰 가중치를 주는 PageRank 유사 방식이다.
SNIP: 학문 분야마다 다른 인용 관행과 인용 가능성을 보정한 논문당 영향력 지표이다.
JIF: 기본적으로 특정 연도에 직전 2년간 게재된 논문 등이 받은 인용을 해당 기간의 인용 가능 항목 수로 나눈 지표이다.
CiteScore: 일정 기간에 출판된 대상 문헌의 피인용 수를 같은 기간의 대상 문헌 수로 나눈 평균 인용 지표이다.
Cited Half-Life: 한 학술지가 받은 인용을 논문의 출판 연도별로 배열했을 때 인용의 절반이 갈리는 논문 연령의 중앙값이다.

NoteBase

선지별 해설

① X — SNIP은 분야별 인용 관행과 인용 가능성을 보정하여 서로 다른 분야의 학술지를 비교하는 지표이다. 틀린 표현은 '5년 동안 출판된 총 논문 수와 피인용횟수를 기반으로 측정한다'이며, 올바르게는 '학술지의 논문당 인용 영향력을 해당 분야의 인용 가능성으로 정규화한다'이다.
② O — Eigenfactor는 최근 5년간의 인용 자료와 학술지 인용 네트워크를 이용해 다른 학술지와 비교되는 상대적 총 영향력을 평가한다. 또한 인용한 학술지의 영향력까지 가중한다.
③ X — 기본 JIF는 직전 2년간 게재된 논문 등이 특정 연도에 받은 인용을 중심으로 계산하는 평균 인용 지표이다. 틀린 표현은 '페이지랭크 알고리즘과 유사한 방식'이며, 올바르게는 '인용 저널의 영향력을 재귀적으로 가중하지 않고 인용 수와 인용 가능 항목 수의 비율로 계산한다'이다.
④ X — CiteScore는 대상 문헌의 평균 피인용도를 계산하는 지표이다. 틀린 표현은 '페이지랭크 알고리즘과 유사한 방식'이며, 올바르게는 '일정 기간의 피인용 수를 같은 기간의 대상 문헌 수로 나누어 계산한다'이다.
⑤ X — Cited Half-Life는 학술지 영향력의 크기가 아니라 인용된 논문들의 시간적 수명을 나타낸다. 틀린 표현은 '상대적 영향력을 평가한다'이며, 올바르게는 '받은 인용의 절반을 기준으로 인용된 논문의 연령 중앙값을 측정한다'이다.

오답 포인트

'피인용횟수'만 보고 JIF나 CiteScore를 고르면 안 된다. 이 문항의 결정적 단서는 '5년', '상대적 영향력', 'PageRank 유사 방식'의 결합이며, 특히 영향력 있는 학술지의 인용에 더 큰 가중치를 주는 네트워크 계산이 Eigenfactor를 가리킨다.

다시 나오면 이렇게 판단

분야별 인용 관행 보정은 SNIP, 기본 2년 평균 인용은 JIF, 일정 기간의 문헌당 평균 인용은 CiteScore, 인용 논문의 연령 중앙값은 Cited Half-Life, 5년 인용 네트워크와 PageRank형 영향력 가중은 Eigenfactor로 판단한다.

암기 포인트

Eigenfactor는 '5년·인용 네트워크·영향력 가중', Cited Half-Life는 '인용 나이의 중앙값'이다.

NoteBase

2. 다음 설명이 지칭하는 에뮬레이션(emulation) 방식을 각각 옳게 나열한 것은?

> ㄱ. 원본에 접근하는 데 요구되는 모든 기술(S/W, H/W, O/S 등)을 보존하는 방식이다.
>
> ㄴ. 디지털 원본에 적용된 기술적인 조건들에 변경이 있어도 인코딩되어 있는 콘텐츠를 재생할 수 있는 환경을 프로그램으로 만들어 내어 디지털 문헌의 접근성을 보장하는 방식이다.
>
> ㄷ. 디지털 파일의 내용 해석을 UVC(Universal Virtual Computer)의 기계언어로 처리할 수 있도록 프로그램을 작성하여 해결하려는 시도이다.

	ㄱ	ㄴ	ㄷ
①	기술보존	기술 에뮬레이션	가상기계 프로그램
②	기술보존	가상기계 프로그램	기술 에뮬레이션
③	가상기계 프로그램	기술 에뮬레이션	기술보존
④	가상기계 프로그램	기술보존	기술 에뮬레이션
⑤	기술 에뮬레이션	기술보존	가상기계 프로그램

정답 1번

출제 포인트

이 문제는 디지털 보존을 위한 에뮬레이션 방식인 기술보존, 가상기계 프로그램, 기술 에뮬레이션의 정의를 구별하는 문제이다.

핵심 개념

기술보존: 디지털 원본에 접근하는 데 필요한 소프트웨어, 하드웨어, 운영체제 등 기술 환경 전체를 함께 보존하는 방식이다. / 키워드: 요구되는 모든 기술 보존

가상기계 프로그램: 기술 환경이 바뀌어도 인코딩된 콘텐츠를 재현할 수 있는 실행 환경을 프로그램으로 구현하는 방식이다. / 키워드: 접근 환경을 프로그램으로 구현

기술 에뮬레이션: 원래 시스템의 기능이나 동작을 다른 시스템에서 흉내 내도록 프로그램을 작성하는 방식이며, UVC는 디지털 파일의 내용을 기계언어 수준에서 처리하는 접근에 해당한다. / 키워드: 원래 시스템의 동작 모방

따라서 ㄱ은 기술보존, ㄴ은 가상기계 프로그램, ㄷ은 기술 에뮬레이션이다.

NoteBase

선지별 해설

① O — ㄱ의 '요구되는 모든 기술을 보존'은 기술보존, ㄴ의 '재생 환경을 프로그램으로 구현'은 가상기계 프로그램, ㄷ의 'UVC의 기계언어로 처리'는 기술 에뮬레이션이므로 세 연결이 모두 옳다.
② X — ㄴ을 기술 에뮬레이션, ㄷ을 가상기계 프로그램으로 연결한 부분이 틀렸다. 올바르게는 ㄴ이 가상기계 프로그램이고 ㄷ이 기술 에뮬레이션이다.
③ X — ㄱ을 가상기계 프로그램, ㄴ을 기술 에뮬레이션, ㄷ을 기술보존으로 연결한 부분이 모두 틀렸다. 올바르게는 ㄱ이 기술보존, ㄴ이 가상기계 프로그램, ㄷ이 기술 에뮬레이션이다.
④ X — ㄱ을 가상기계 프로그램, ㄴ을 기술보존, ㄷ을 기술 에뮬레이션으로 연결한 부분 중 ㄱ과 ㄴ이 틀렸다. 올바르게는 ㄱ이 기술보존이고 ㄴ이 가상기계 프로그램이며, ㄷ의 기술 에뮬레이션 연결만 옳다.
⑤ X — ㄱ을 기술 에뮬레이션, ㄴ을 기술보존, ㄷ을 가상기계 프로그램으로 연결한 부분이 모두 틀렸다. 올바르게는 ㄱ이 기술보존, ㄴ이 가상기계 프로그램, ㄷ이 기술 에뮬레이션이다.

오답 포인트

'프로그램으로 만들어 낸다'는 표현만 보고 ㄴ과 ㄷ을 같은 방식으로 판단하기 쉽다. ㄴ은 변화한 기술 조건에서도 콘텐츠를 재생할 수 있는 환경 자체를 프로그램으로 구현한다는 점이 핵심이고, ㄷ은 UVC의 기계언어로 디지털 파일의 내용을 처리한다는 점이 핵심이다.

다시 나오면 이렇게 판단

'필요한 모든 기술 환경을 그대로 보존'하면 기술보존, '재생 가능한 환경을 프로그램으로 구현'하면 가상기계 프로그램, '원래 시스템의 기능을 다른 환경에서 모방하거나 UVC로 처리'하면 기술 에뮬레이션으로 판단한다.

암기 포인트

기술보존은 환경을 보존하고, 가상기계는 환경을 구현하며, 기술 에뮬레이션은 동작을 모방한다.

NoteBase

3. 자연어 색인의 장점으로 옳은 것만을 <보기>에서 모두 고르면?

< 보 기 >

ㄱ. 다양한 접근점의 제공이 가능하다.
ㄴ. 새로운 개념의 표현이 용이하다.
ㄷ. 용어의 계층관계 표시가 수월하다.
ㄹ. 융통성과 표현력이 좋다.
ㅁ. 동의어, 동형이의어의 통제가 쉽다.

① ㄱ, ㄴ
② ㄷ, ㅁ
③ ㄱ, ㄴ, ㄹ
④ ㄷ, ㄹ, ㅁ
⑤ ㄱ, ㄴ, ㄷ, ㅁ

정답 3번

출제 포인트

자연어 색인의 장점과 단점을 통제어 색인과 대비하여 구분하는 문제이다.

핵심 개념

자연어 색인은 문헌이나 이용자의 표현을 그대로 색인어로 사용하는 방식이다. 어휘 선택의 제약이 적어 다양한 접근점을 제공하고, 새로운 개념을 빠르게 표현하며, 융통성과 표현력이 좋다. 반면 어휘를 통제하지 않으므로 동의어·동형이의어의 통제가 어렵고 용어 사이의 계층관계를 일관되게 나타내기 어렵다.
통제어 색인은 시소러스 등에서 미리 정한 용어를 사용하여 동의어를 통일하고 계층관계를 명시하는 방식이다.

선지별 해설

① X — ㄱ과 ㄴ은 장점이지만 ㄹ도 장점이므로 불완전하다. 올바른 조합은 ㄱ, ㄴ, ㄹ이다.
② X — ㄷ의 '계층관계 표시가 수월하다'와 ㅁ의 '동의어, 동형이의어의 통제가 쉽다'가 틀렸다. 자연어 색인에서는 계층관계 표시와 동의어·동형이의어 통제가 어렵다. 올바른 조합은 ㄱ, ㄴ, ㄹ이다.
③ O — ㄱ의 다양한 접근점, ㄴ의 새로운 개념 표현 용이성, ㄹ의 높은 융통성과 표현력은 모두 자연어 색인의 장점이다.
④ X — ㄹ은 장점이지만 ㄷ과 ㅁ은 장점이 아니다. 자연어 색인에서는 용어의 계층관계 표시와 동의어·동형이의어 통제가 어렵다. 올바른 조합은 ㄱ, ㄴ, ㄹ이다.
⑤ X — ㄱ과 ㄴ은 장점이지만 ㄷ과 ㅁ은 단점이며, 장점인 ㄹ이 빠져 있다. ㄷ과 ㅁ은 각각 '계층관계 표시가 어렵다', '동의어·동형이의어 통제가 어렵다'로 고쳐야 하며, 올바른 조합은 ㄱ, ㄴ, ㄹ이다.

NoteBase

오답 포인트

자연어의 '표현이 자유롭다'는 장점과 '용어 통제가 잘된다'를 같은 뜻으로 혼동하기 쉽다. 표현의 자유는 ㄱ·ㄴ·ㄹ로 나타나지만, 그 자유 때문에 오히려 ㄷ의 계층관계 표시와 ㅁ의 동의어·동형이의어 통제는 어려워진다.

다시 나오면 이렇게 판단

자연어 색인은 자유로운 표현·신속한 신개념 반영·다양한 접근점을 장점으로 판단한다. 용어의 통일, 동의어·동형이의어 통제, 계층관계 명시는 통제어 색인의 장점으로 구분한다.

암기 포인트

자연어는 표현은 자유롭고 통제는 어렵다.

NoteBase

4. 다음은 연구자 A의 논문을 피인용횟수가 많은 순으로 정렬한 표이다. 연구자 A의 h지수는?

논문명	피인용순위	피인용횟수
가	1	45
나	2	30
다	3	24
라	4	15
마	5	14
바	6	11
사	7	9
아	8	7
자	9	6
차	10	5

① 5
② 6
③ 7
④ 8
⑤ 9

정답 3번

출제 포인트

피인용횟수를 내림차순으로 놓았을 때, 순위 이상으로 인용된 논문의 최대 개수를 이용해 h지수를 구하는 문제이다.

조건 정리

주어진 값: 순위별 피인용횟수는 45, 30, 24, 15, 14, 11, 9, 7, 6, 5이다.
구해야 할 값: 연구자 A의 h지수이다.
주의할 조건: 단순히 조건을 만족하는 순위를 찾는 것이 아니라, 그중 가장 큰 순위를 찾아야 한다.

사용할 공식

피인용순위가 r인 논문의 피인용횟수를 c_r라고 하면,

$$h = \max\{r \mid c_r \geq r\}$$

를 사용한다. h지수는 피인용횟수가 각각 h회 이상인 논문을 h편 이상 보유할 때 가능한 가장 큰 h이기 때문이다.

NoteBase

계산 과정

1. 7위 논문의 피인용횟수는 9회이므로 c_7=9≥7이다. 따라서 적어도 7편의 논문이 각각 7회 이상 인용되었다.
2. 8위 논문의 피인용횟수는 7회이므로 $c_8 = 7 < 8$이다. 따라서 8편의 논문이 각각 8회 이상 인용되었다는 조건은 성립하지 않는다.
3. 조건을 만족하는 최대 순위가 7이므로 h지수는 7이다.

선지 판단

① X — 5위 논문의 피인용횟수 14회는 5회 이상이지만, 5는 조건을 만족하는 최대값이 아니다. 7위까지 조건이 성립하므로 h지수는 5가 아니라 7이다.
② X — 6위 논문의 피인용횟수 11회는 6회 이상이지만, 7위도 9회로 7회 이상 인용되었다. 따라서 h지수는 6이 아니라 7이다.
③ O — 7위 논문이 9회 인용되어 7편 모두 7회 이상 인용되었고, 8위에서는 조건이 깨지므로 h지수는 7이다.
④ X — 8위 논문의 피인용횟수는 7회로 8회에 미치지 못한다. 따라서 h지수는 8이 아니라 7이다.
⑤ X — 9위 논문의 피인용횟수는 6회로 9회에 미치지 못하고, 이미 8위에서도 조건이 성립하지 않는다. 따라서 h지수는 9가 아니라 7이다.

오답 포인트

①과 ②처럼 조건을 만족하는 아무 순위에서 계산을 멈추거나, ④처럼 8위 논문의 피인용횟수 7을 순위 8과 비교하지 않고 논문 수만 세면 틀리기 쉽다. h지수는 조건을 만족하는 최대 순위이며, 처음으로 조건이 깨지는 순위의 바로 앞값을 확인해야 한다.

다시 나오면 이렇게 판단

피인용횟수를 내림차순으로 정렬한 뒤 각 순위 r와 해당 피인용횟수 c_r를 비교한다. $c_r \geq r$가 마지막으로 성립하는 순위를 h지수로 판단한다.

암기 포인트

h지수는 ‘h회 이상 인용된 논문이 h편 이상’이 되는 가장 큰 h이다.

NoteBase

5. HTML과 XML에 대한 설명으로 옳은 것만을 <보기>에서 모두 고르면?

< 보 기 >

ㄱ. HTML은 W3C(World Wide Web Consortium) 표준이다.

ㄴ. HTML은 이종의 컴퓨터들이 처리할 수 있는 웹 문서를 작성하는 언어이다.

ㄷ. HTML과 XML 모두 태그(tag)를 확장하여 사용 가능하다.

ㄹ. XML은 국제표준이며 상이한 응용 프로그램 간의 호환을 가능하게 한다.

ㅁ. HTML은 데이터 구조를 기술하는 기능에 한계가 없다.

① ㄴ, ㄹ

② ㄱ, ㄴ, ㄹ

③ ㄱ, ㄴ, ㅁ

④ ㄷ, ㄹ, ㅁ

⑤ ㄱ, ㄴ, ㄹ, ㅁ

정답 2번

출제 포인트

HTML과 XML의 표준성, 용도, 태그 확장성, 호환성 차이를 구분하는 문제이다.

핵심 개념

HTML: 웹 문서의 구조와 표현을 기술하는 표준 마크업 언어이며, 정해진 태그를 사용한다.
XML: 데이터를 구조화하여 표현하고 교환하기 위한 확장 가능한 표준 마크업 언어이며, 사용자가 목적에 맞는 태그를 정의할 수 있다.
핵심 차이는 태그 확장성이다. XML은 태그를 정의하여 확장할 수 있지만 HTML의 태그 집합은 정해져 있다.
따라서 옳은 진술은 ㄱ, ㄴ, ㄹ이다.

NoteBase

선지별 해설

① X — ㄴ과 ㄹ만 고른 조합은 옳은 ㄱ을 빠뜨렸다. ㄱ의 HTML은 웹 문서 작성을 위한 표준 마크업 언어라는 설명도 옳으므로, 올바른 조합은 ㄱ, ㄴ, ㄹ이다.
② O — ㄱ은 HTML의 표준성, ㄴ은 서로 다른 컴퓨터가 처리할 수 있는 웹 문서 작성 언어라는 성격, ㄹ은 XML의 표준성과 응용 프로그램 간 데이터 교환·호환 지원을 바르게 설명한다.
③ X — ㄱ과 ㄴ은 옳지만 ㅁ의 '데이터 구조를 기술하는 기능에 한계가 없다'는 표현이 틀렸다. HTML은 정해진 태그를 사용하므로 임의의 데이터 구조를 표현하는 데 한계가 있으며, 확장 가능한 데이터 구조 기술에 적합한 것은 XML이다. 올바른 조합은 ㄱ, ㄴ, ㄹ이다.
④ X — ㄹ은 옳지만 ㄷ과 ㅁ이 틀렸다. ㄷ은 'HTML과 XML 모두'가 아니라 XML만 사용자가 태그를 정의하여 확장할 수 있다고 고쳐야 한다. ㅁ은 HTML에 데이터 구조 기술의 한계가 있다고 고쳐야 한다. 올바른 조합은 ㄱ, ㄴ, ㄹ이다.
⑤ X — ㄱ, ㄴ, ㄹ은 옳지만 ㅁ까지 포함한 것이 틀렸다. ㅁ의 '한계가 없다'는 표현을 '정해진 태그 때문에 임의의 데이터 구조 기술에는 한계가 있다'로 고쳐야 한다. 올바른 조합은 ㄱ, ㄴ, ㄹ이다.

오답 포인트

ㄷ에서 XML의 확장성을 HTML에도 그대로 적용한 점과, ㅁ에서 HTML의 웹 문서 작성 기능을 데이터 구조의 무제한 기술 능력으로 과장한 점이 함정이다.

다시 나오면 이렇게 판단

'사용자가 태그를 정의한다', '데이터 구조를 유연하게 표현한다', '응용 프로그램 간 데이터를 교환한다'는 표현은 XML의 특징으로 판단한다. HTML은 정해진 태그로 웹 문서의 구조와 표현을 기술한다.

암기 포인트

HTML은 정해진 태그로 웹 문서를 작성하고, XML은 정의한 태그로 데이터를 구조화·교환한다.

NoteBase

6. 불리언 모델에서 문헌의 색인어뿐만 아니라 질의어에도 가중치를 부여하는 방법으로 옳은 것은?

① 퍼지집합 모델
② MMM 모델
③ P-Norm 모델
④ 필드 가중치 기법
⑤ 논리합 정규형 기법

정답 3번

출제 포인트

불리언 검색의 엄격한 일치 방식을 완화하는 여러 확장 모델 가운데, 문헌의 색인어와 질의어 양쪽에 가중치를 부여하는 모델을 구별하는 문제이다.

핵심 개념

P-Norm 모델은 불리언 질의의 AND·OR 구조를 유지하면서 문헌의 색인어와 질의어에 가중치를 부여하고, p-노름을 이용해 문헌과 질의의 유사도를 연속적인 값으로 계산하는 확장 불리언 모델이다. 따라서 질의어의 상대적 중요도까지 검색 결과에 반영할 수 있다.
퍼지집합 모델과 MMM 모델도 엄격한 이진 판정을 완화하지만, 이 문항에서 묻는 '문헌 색인어와 질의어 모두의 가중치'라는 결정적 특징은 P-Norm 모델에 해당한다. 필드 가중치 기법은 제목·초록·본문과 같은 문헌 필드별 중요도를 다르게 주는 방식이다.

NoteBase

선지별 해설

① X — 퍼지집합 모델은 문헌이 질의어에 어느 정도 소속되는지를 연속적인 소속도 값으로 표현하여 불리언 검색을 완화하는 모델이다. 틀린 표현은 '질의어에도 가중치를 부여하는 방법'으로 본 부분이며, 올바르게는 문헌 색인어와 질의어 양쪽의 가중치를 명시적으로 반영하는 모델은 P-Norm 모델이다.
② X — MMM 모델은 AND의 최소 연산과 OR의 최대 연산을 혼합·조정하여 불리언 연산의 엄격성을 완화하는 모델이다. 틀린 표현은 이를 '문헌 색인어와 질의어 모두에 가중치를 부여하는 방법'으로 본 부분이며, 올바르게는 그 특징에 해당하는 모델은 P-Norm 모델이다.
③ O — P-Norm 모델은 문헌의 색인어 가중치뿐 아니라 질의어의 중요도를 나타내는 가중치도 반영할 수 있으므로 문제의 설명과 일치한다.
④ X — 필드 가중치 기법은 제목·초록·본문 등 문헌을 구성하는 필드마다 검색상의 중요도를 다르게 부여하는 기법이다. 틀린 표현은 '질의어에도 가중치를 부여한다'는 부분이며, 올바르게는 문헌의 필드별 가중치를 조정하는 기법이다.
⑤ X — 논리합 정규형 기법은 불리언 질의를 OR로 연결된 여러 AND 항의 형태로 정규화하여 처리하는 표현·변환 방식이다. 틀린 표현은 이를 '문헌과 질의어 양쪽의 가중치 부여 모델'로 본 부분이며, 올바르게는 질의의 논리 구조를 정규형으로 바꾸는 기법이다.

오답 포인트

퍼지집합 모델과 MMM 모델도 불리언 모델의 이진적 판단을 완화하므로 P-Norm 모델과 혼동하기 쉽다. 그러나 이 문항의 핵심 표현은 '질의어에도 가중치'이며, 이 표현을 P-Norm 모델과 연결해야 한다. 필드 가중치는 질의어가 아니라 문헌 내부의 필드에 부여하는 가중치이다.

다시 나오면 이렇게 판단

선지의 명칭보다 가중치가 부여되는 대상을 먼저 확인한다. 문헌 색인어와 질의어 양쪽이면 P-Norm 모델, 제목·초록·본문 같은 문헌 영역이면 필드 가중치 기법으로 판단한다.

암기 포인트

P-Norm은 문헌 가중치와 질의 가중치를 함께 반영한다. 필드 가중치는 문헌의 영역별 중요도를 반영한다.

NoteBase

7. 다음 CCL(Creative Commons License)의 이용허락조건에 대한 설명으로 옳지 않은 것은?



① ㄱ은 '저작자 표시'로 저작물을 복사하거나 다른 곳에 게시할 때도 반드시 저작자와 출처를 표시해야 함을 의미한다.

② ㄴ은 '비영리'를 표시하며 저작물을 영리목적으로 이용하려면 별도의 계약이 필요함을 의미한다.

③ ㄷ은 '2차 저작물 표시'로 원 저작물을 변경한 2차 저작물일 경우에 기입해야 함을 의미한다.

④ ㄹ은 '동일조건변경허락'으로 2차 저작물 창작을 허용하되, 원 저작물과 동일한 라이선스를 적용해야 함을 의미한다.

⑤ 이용허락조건 4종류를 조합하여 총 6종류의 CCL이 존재한다.

정답 3번

출제 포인트

CCL의 네 가지 이용허락조건인 저작자 표시(BY), 비영리(NC), 변경금지(ND), 동일조건변경허락(SA)의 기호와 의미를 구별하는 문제이다.

핵심 개념

CCL은 저작자가 자신의 저작물에 적용할 이용 조건을 미리 표시하여, 이용자가 그 조건을 지키는 범위에서 저작물을 사용할 수 있게 하는 표준 이용허락 방식이다.
저작자 표시(BY): 저작자와 출처를 표시해야 한다.
비영리(NC): 영리 목적으로 이용할 수 없다.
변경금지(ND): 저작물을 변경하거나 2차적 저작물로 만들어 이용할 수 없다.
동일조건변경허락(SA): 변경·2차적 저작물 작성을 허용하되 원 저작물과 동일한 라이선스를 적용해야 한다.

NoteBase

선지별 해설

① O — ㄱ의 사람 모양 기호는 저작자 표시(BY)이다. 복제하거나 게시하는 경우에도 저작자와 출처를 표시해야 한다.
② O — ㄴ의 달러 기호에 빗금이 있는 표시는 비영리(NC)이다. 영리 목적으로 이용하려면 해당 CCL의 허용 범위를 벗어나므로 저작권자의 별도 허락이 필요하다.
③ X — 틀린 표현은 ㄷ을 '2차 저작물 표시'라고 한 부분이다. ㄷ의 등호 기호는 변경금지(ND)이며, 올바르게는 저작물의 변경과 2차적 저작물 작성을 허용하지 않는다는 뜻이다.
④ O — ㄹ의 원형 화살표 기호는 동일조건변경허락(SA)이다. 2차적 저작물의 창작은 허용하지만 그 결과물에도 원 저작물과 동일한 라이선스를 적용해야 한다.
⑤ O — 저작자 표시, 비영리, 변경금지, 동일조건변경허락의 네 조건을 허용 가능한 방식으로 조합하면 저작자 표시, 저작자 표시-비영리, 저작자 표시-변경금지, 저작자 표시-동일조건변경허락, 저작자 표시-비영리-변경금지, 저작자 표시-비영리-동일조건변경허락의 6종류 CCL이 된다.

오답 포인트

③은 2차적 저작물과 관련된 조건이라는 점만 보고 의미를 반대로 판단하게 만든다. 등호(ND)는 2차적 저작물임을 표시하라는 뜻이 아니라 변경 자체를 금지한다는 뜻이다. 반면 원형 화살표(SA)는 변경을 허용하되 같은 조건을 이어서 적용한다.

다시 나오면 이렇게 판단

기호의 생김새를 핵심 동작과 연결한다. 사람은 저작자 표시, 달러 빗금은 비영리, 등호는 변경 없이 그대로, 원형 화살표는 같은 라이선스의 계승으로 판단한다.

암기 포인트

BY는 표시, NC는 비영리, ND는 변경금지, SA는 동일조건 계승이다.

NoteBase

8. BIBFRAME에 대한 설명으로 옳지 않은 것은?

① 핵심 상위 클래스인 인스턴스는 저자, 언어, 주제를 포함한다.
② 서지정보를 시멘틱웹에서 사용할 수 있도록 링크드데이터 형식으로 구축하고 RDF/XML로 인코딩한다.
③ 부가적 클래스인 이벤트는 저작의 내용과 관련된 사건이다.
④ 개별자료 클래스는 인스턴스의 실질적 사본의 물리적 혹은 가상의 소장위치를 포함한다.
⑤ 부가적 클래스인 에이전트는 개인, 단체, 가족, 정부 등을 포함한다.

정답 1번

출제 포인트

이 문제는 BIBFRAME의 핵심 클래스인 Work, Instance, Item과 부가 클래스인 Event, Agent의 의미를 구분하는 문제이다.

핵심 개념

Work(저작): 자료의 지적·개념적 내용을 나타내며 저자, 언어, 주제 같은 정보를 담는다.
Instance(인스턴스): Work가 특정 출판·발행 형태로 구현된 판이나 형식을 나타낸다.
Item(개별자료): 특정 Instance의 실제 개별 사본이며 소장 위치 등의 정보를 나타낸다.
Event(이벤트): 저작의 내용과 관련된 사건을 나타내는 부가 클래스이다.
Agent(에이전트): 저작이나 자원과 관련된 개인, 단체, 가족, 관할기관 등의 행위 주체를 나타내는 부가 클래스이다.

선지별 해설

① X — '인스턴스는 저자, 언어, 주제를 포함한다'는 부분이 틀렸다. 저자, 언어, 주제는 저작의 지적·개념적 내용을 설명하는 Work에 해당한다. 올바르게는 '핵심 상위 클래스인 Work는 저자, 언어, 주제를 포함한다'이다.
② O — BIBFRAME은 서지정보를 웹의 연결 데이터로 활용할 수 있도록 링크드데이터 방식으로 구축하며, RDF를 기반으로 표현할 수 있다. RDF/XML은 그 인코딩 형식 가운데 하나이다.
③ O — Event는 저작의 내용과 관련된 사건을 표현하는 부가 클래스이다.
④ O — Item은 Instance의 실제 개별 사본으로서 물리적 또는 전자적 소장 대상과 그 소장 위치를 표현한다.
⑤ O — Agent는 자원과 관련된 행위 주체를 나타내며 개인, 단체, 가족, 정부·관할기관 등을 포함한다.

오답 포인트

①은 Work와 Instance의 역할을 바꾸어 제시했다. '저자·언어·주제'는 내용의 정체성을 설명하므로 Work에 속하고, Instance는 그 내용이 어떤 판이나 발행 형태로 구현되었는지를 나타낸다.

NoteBase

다시 나오면 이렇게 판단

내용의 지적 정체성은 Work, 출판·발행된 구현형은 Instance, 실제 소장 사본은 Item으로 판단한다. 사건은 Event, 사람이나 단체 같은 행위 주체는 Agent로 구분한다.

암기 포인트

Work는 내용, Instance는 구현, Item은 사본이다. Event는 사건, Agent는 행위 주체이다.

NoteBase

9. 다음 문헌분리값 공식에 기반한 설명으로 옳지 않은 것은?

> **문헌분리값$_k$ = Q_k − Q**
>
> 문헌분리값$_k$: 단어 k의 문헌분리값
> Q_k : 단어 k를 제거한 후의 평균 유사도
> Q : 단어 k를 색인어로 부여하였을 때 평균 유사도

① 단어의 문헌분리값은 특정한 단어가 한 문헌집단에서 상호 관련없는 문헌을 분리하는 능력을 측정한 것이다.

② 문헌분리값이 큰 단어일수록 좋은 색인어가 된다.

③ 단어 k가 주제가 다른 문헌을 가능한 한 분리시키는 좋은 색인어라면 $Q_k - Q$는 양수가 된다.

④ 단어 k가 많은 문헌에 출현하는 경우 $Q_k - Q$는 음수가 될 수 있다.

⑤ 단어 k의 가중치는 문헌빈도와 문헌분리값을 곱하여 계산한다.

정답 5번

출제 포인트

문헌분리값의 정의와 부호가 나타내는 색인어의 분리 능력을 이해하고, 단어 가중치에 사용되는 빈도의 종류를 구분하는 문제이다.

조건 정리

Q_k는 단어 k를 제거한 뒤의 평균 유사도이고, Q는 단어 k를 색인어로 부여했을 때의 평균 유사도이다. 좋은 색인어는 관련 없는 문헌들을 서로 구별하므로, 그 단어를 포함했을 때의 평균 유사도 Q가 제거했을 때의 평균 유사도 Q_k보다 작아진다.

사용할 공식

문헌분리값은 다음과 같다.

$$D_k = Q_k - Q$$

따라서 $D_k > 0$이면 단어 k가 문헌을 잘 분리하는 좋은 색인어이고, 값이 클수록 분리 능력이 크다. 반대로 $D_k < 0$이면 그 단어가 문헌들을 불필요하게 서로 비슷하게 만드는 경향이 있음을 뜻한다.

NoteBase

선지 판단

① O — 문헌분리값은 특정 단어가 문헌집단에서 서로 관련 없는 문헌을 구별하는 능력을 나타내므로 옳다.

② O — 문헌분리값이 크다는 것은 그 단어를 색인어로 사용했을 때 평균 유사도가 더 크게 낮아져 문헌의 구분이 잘 된다는 뜻이므로 옳다.

③ O — 좋은 색인어라면 그 단어를 포함한 경우의 평균 유사도 Q가 제거한 경우의 평균 유사도 Q_k보다 작으므로 $Q_k - Q > 0$이 된다.

④ O — 많은 문헌에 공통으로 나타나는 단어는 문헌을 구별하기보다 서로 비슷하게 보이게 할 수 있다. 이 경우 $Q > Q_k$가 되어 $Q_k - Q < 0$이 될 수 있다.

⑤ X — '문헌빈도와 문헌분리값을 곱한다'는 표현이 틀렸다. 올바르게는 단어 k의 가중치는 해당 문헌에서의 단어빈도와 문헌분리값을 곱하여 계산한다. 문헌빈도는 그 단어가 나타난 문헌의 수이므로 단어빈도와 구별해야 한다.

오답 포인트

⑤는 '단어빈도'를 '문헌빈도'로 바꾼 것이 함정이다. 문헌분리값이 색인어의 전역적인 분리 능력을 나타낸다는 점만 기억하고 빈도 용어를 확인하지 않으면 놓치기 쉽다. 또한 공식의 차례가 $Q_k - Q$이므로 좋은 색인어의 문헌분리값은 양수라는 점도 함께 구분해야 한다.

다시 나오면 이렇게 판단

먼저 Q_k가 '단어를 제거한 경우'이고 Q가 '단어를 사용한 경우'임을 확인한다. 이어 좋은 색인어는 사용했을 때 문헌 사이 평균 유사도를 낮추므로 $Q_k - Q$가 양수라고 판단한다. 가중치에 관한 선지는 문헌빈도가 아니라 해당 문헌에서의 단어빈도를 사용하는지 확인한다.

암기 포인트

분리값은 '제거 후 유사도 − 사용 시 유사도'이고, 클수록 좋은 색인어이다. 단어 가중치는 '단어빈도 × 문헌분리값'이다.

NoteBase

10. 텍스트 마이닝에 대한 설명으로 옳지 않은 것은?

① 자연어 문서에 수록된 비정형 텍스트 데이터를 분석하여 어떤 규칙이나 패턴 등 유용한 정보를 찾는 것이다.

② 대소문자 통일, 문장부호·특수문자 제거, 불용어 삭제 등의 데이터 전처리 과정이 필요하다.

③ 감성분석은 단어와 문맥을 분석하여 텍스트에 들어 있는 사람들의 태도, 의견, 성향과 같은 주관적인 데이터를 분석하는 기술이다.

④ 단어의 출현빈도를 분석하는 단어 구름(word cloud)은 핵심 단어를 추출하여 문서의 주제를 객관적이면서 명확하게 식별하는 것이다.

⑤ 주요 응용분야로는 정보 추출, 문서 요약, 문서 분류 등이 있으며, 부가적으로 분석결과에 대한 정보 시각화를 수행하기도 한다.

정답 4번

출제 포인트

텍스트 마이닝의 개념, 전처리 과정, 주요 분석 기법과 활용 분야를 구분하고, 워드 클라우드가 제공하는 정보의 한계를 판단하는 문제이다.

핵심 개념

텍스트 마이닝은 자연어 문서와 같은 비정형 텍스트에서 패턴과 의미 있는 정보를 추출하는 분석 과정이다. 분석 전에는 대소문자 통일, 문장부호·특수문자 제거, 불용어 삭제 등의 전처리를 수행할 수 있다. 감성분석은 텍스트에 나타난 태도·의견·성향을 분석하고, 워드 클라우드는 단어의 출현빈도 등을 글자 크기로 시각화한다. 워드 클라우드는 빈도 분포를 직관적으로 보여 주지만 문맥과 단어 간 관계까지 설명하지는 않는다.

NoteBase

선지별 해설

① O — 텍스트 마이닝은 자연어 문서에 담긴 비정형 텍스트를 분석하여 규칙, 패턴, 관계 등 유용한 정보를 찾아내는 기술이다.
② O — 표기 차이와 분석 잡음을 줄이기 위해 대소문자를 통일하고 문장부호·특수문자와 불용어를 제거하는 전처리를 수행할 수 있다.
③ O — 감성분석은 단어와 문맥을 바탕으로 텍스트에 나타난 사람의 태도, 의견, 성향과 같은 주관적 정보를 분석하는 기술이다.
④ X — 틀린 부분은 워드 클라우드가 핵심 단어를 추출하여 문서의 주제를 '객관적이면서 명확하게' 식별한다는 표현이다. 워드 클라우드는 주로 단어의 출현빈도나 가중치를 시각적으로 나타내는 도구이며, 문맥과 단어 간 관계를 충분히 반영하지 못하므로 문서 주제를 객관적·명확하게 확정하는 기법이라고 할 수 없다. 올바르게는 '워드 클라우드는 단어 빈도나 가중치를 시각화하여 자주 등장하거나 중요도가 높게 설정된 단어를 직관적으로 파악하게 한다'이다.
⑤ O — 텍스트 마이닝은 정보 추출, 문서 요약, 문서 분류 등에 활용되며, 분석 결과를 그래프나 워드 클라우드 등으로 시각화하기도 한다.

오답 포인트

④는 '자주 나온 단어를 크게 보여 준다'는 워드 클라우드의 기능을 '문서의 주제를 객관적이고 명확하게 식별한다'는 기능으로 과장한 선지이다. 단어 빈도가 높아도 문맥상 핵심 주제와 반드시 일치하는 것은 아니며, 같은 빈도라도 단어 사이의 관계나 의미는 드러나지 않는다.

다시 나오면 이렇게 판단

텍스트 마이닝 선지는 '전처리—분석—활용'으로 나누어 판단한다. 워드 클라우드는 빈도·가중치의 시각화 도구이고, 주제의 객관적 확정이나 문맥 분석 도구는 아니라는 기준을 적용한다.

암기 포인트

텍스트 마이닝은 비정형 텍스트에서 정보를 찾고, 감성분석은 태도·의견을 읽으며, 워드 클라우드는 단어 빈도를 보여 준다. '빈도 시각화'와 '주제의 객관적 식별'은 다르다.

NoteBase

11. 다음에서 설명하는 메타데이터 표준에 정의된 요소명이 아닌 것은?

> ○ MARC, ONIX, TEI 등을 절충하여 상호운용성과 정밀성을 만족시키도록 설계되었다.
> ○ 디지털 도서관 영역의 범용 서지정보 메타데이터로 MARC와 상호호환이 가능하다.
> ○ XML 스키마 기반의 인코딩 구조를 채택하며 언어형식의 태그를 사용한다.

① typeOfResource
② accessCondition
③ identifier
④ title
⑤ originInfo

정답 4번

출제 포인트

이 문제는 MODS(Metadata Object Description Schema)의 최상위 요소명을 구별하는 문제이다.

핵심 개념

MODS는 도서관 서지정보를 XML로 표현하는 메타데이터 표준으로, MARC와의 호환성을 고려하면서 MARC보다 간명한 요소 체계를 사용한다. 제목 정보의 최상위 요소명은 title이 아니라 titleInfo이며, 실제 제목은 titleInfo 내부의 title 하위 요소에 기록한다.

선지별 해설

① O — typeOfResource는 자료의 기본 유형을 나타내는 MODS의 최상위 요소이다. 문자 자료, 지도, 악보 등 자원의 일반적 유형을 기술한다.
② O — accessCondition은 자료 이용에 관한 제한이나 이용·재생산 조건을 기술하는 MODS의 최상위 요소이다.
③ O — identifier는 ISBN, ISSN 등 자원을 식별하는 번호나 코드를 기록하는 MODS의 최상위 요소이다.
④ X — 틀린 표현은 title이다. MODS에서 제목 정보를 담는 최상위 요소명은 titleInfo이며, title은 그 안에서 실제 제목을 기록하는 하위 요소이다. 따라서 이 문맥에서는 title을 titleInfo로 고쳐야 한다.
⑤ O — originInfo는 발행지, 발행처, 발행일 등 자료의 생성·발행 정보를 기술하는 MODS의 최상위 요소이다.

NoteBase

오답 포인트

④의 title은 제목을 뜻하는 자연스러운 단어이고 MODS 내부의 하위 요소로도 사용되므로 최상위 요소명처럼 보이기 쉽다. 이 문제에서는 최상위 요소명인 titleInfo를 묻는다는 점이 함정이다.

다시 나오면 이렇게 판단

MODS 요소명 문제에서는 기능을 나타내는 묶음 요소의 정확한 표기를 확인한다. 제목은 titleInfo, 발행 정보는 originInfo, 자료 유형은 typeOfResource로 대응시키고, title처럼 묶음 안에 들어가는 하위 요소와 구별한다.

암기 포인트

MODS의 제목 묶음은 titleInfo이고, 실제 제목값은 그 안의 title이다.

NoteBase

12. OAK(Open Access Korea) 리포지터리에 대한 설명으로 옳은 것만을 <보기>에서 고르면 모두 몇 개인가?

< 보 기 >

ㄱ. 한국형 오픈액세스 리포지터리로 Google 등 해외 검색포털 등과의 연계가 불가하다.

ㄴ. 국제적으로 많이 사용하는 DSpace를 기반으로 개발되었다.

ㄷ. 공공기관, 대학, 연구소 등의 국내 기관에서 생산한 디지털 지식정보를 보존 및 관리한다.

ㄹ. 다양한 포맷(PDF, HWP, DOC 등)의 텍스트 자료만을 아카이빙하고 있다.

ㅁ. 현재 국립중앙도서관이 국가지식 정보 구축 및 확산 사업으로 추진하고 있다.

① 1개
② 2개
③ 3개
④ 4개
⑤ 5개

정답 3번

출제 포인트

OAK 리포지터리의 기반 소프트웨어, 수집·보존 대상, 검색 연계, 운영 주체를 구분하고 옳은 보기의 수를 세는 문제이다.

핵심 개념

OAK 리포지터리는 국내 기관이 생산한 디지털 지식정보를 수집·보존·관리하고 개방하는 한국형 오픈액세스 리포지터리이다.

ㄱ X: ‘Google 등 해외 검색포털과의 연계가 불가능하다’가 틀리다. 실제로는 검색포털과 연계하여 콘텐츠의 검색과 확산이 가능하다.

ㄴ O: 국제적으로 널리 쓰이는 오픈소스 리포지터리 소프트웨어 DSpace를 기반으로 개발되었다.

ㄷ O: 공공기관·대학·연구소 등 국내 기관이 생산한 디지털 지식정보를 보존하고 관리한다.

ㄹ X: ‘텍스트 자료만’이 틀리다. 실제로는 특정 텍스트 포맷에만 한정하지 않고 다양한 유형의 디지털 지식정보를 아카이빙한다.

ㅁ O: 이 문항의 설명에서 OAK는 국립중앙도서관이 국가 지식정보의 구축과 확산을 위해 추진하는 사업이다.

따라서 옳은 보기는 ㄴ, ㄷ, ㅁ의 3개이다.

NoteBase

선지별 해설

① X: 옳은 보기는 1개가 아니라 ㄴ, ㄷ, ㅁ의 3개이다. '1개'는 '3개'로 고쳐야 한다.
② X: 옳은 보기는 2개가 아니라 ㄴ, ㄷ, ㅁ의 3개이다. '2개'는 '3개'로 고쳐야 한다.
③ O: ㄴ, ㄷ, ㅁ이 옳으므로 옳은 보기의 수는 3개이다.
④ X: ㄱ과 ㄹ이 틀리므로 옳은 보기는 4개가 아니라 3개이다. '4개'는 '3개'로 고쳐야 한다.
⑤ X: ㄱ과 ㄹ이 틀리므로 옳은 보기는 5개가 아니라 3개이다. '5개'는 '3개'로 고쳐야 한다.

오답 포인트

ㄱ의 '연계가 불가능하다'와 ㄹ의 '텍스트 자료만'은 범위를 지나치게 제한한 표현이다. 이 두 문장을 맞는 설명으로 처리하면 정답 개수가 달라진다.

다시 나오면 이렇게 판단

리포지터리 문항에서는 기반 소프트웨어, 운영 주체, 수집 대상, 외부 검색 연계를 각각 나누어 판단한다. '불가능하다', '~만'처럼 기능이나 대상을 단정적으로 제한하는 표현은 실제 개방·확산 목적과 맞는지 확인한다.

암기 포인트

OAK는 DSpace 기반이며, 국내 기관의 다양한 디지털 지식정보를 보존·관리하고 검색·확산을 지원한다.

NoteBase

13. 정보검색 모델에 대한 설명으로 옳지 않은 것은?

① LSI(Latent Semantic Indexing) 모델은 대수적(algebraic) 모델에 속하는 개념색인 기법이다.

② 확장 불리언 모델은 용어의 관계를 설정할 수 있지만 부분일치 검색은 허용하지 않는다.

③ 벡터공간 모델은 대수적(algebraic) 모델로 부분일치 검색과 순위화가 가능하다.

④ 확률 모델은 베이즈 정리(Bayes' theorem)를 변형하여 적합성 확률을 구할 수 있다.

⑤ 퍼지집합 모델은 한 문헌이 특정한 색인어로 표현되는 개념 범주에 속하는 정도를 나타내는 부분 멤버십을 이용한다.

정답 2번

출제 포인트

정보검색 모델별 표현 방식과 부분일치 검색·순위화 가능 여부를 구별하는 문제이다.

핵심 개념

LSI: 용어-문헌 행렬을 특잇값 분해하여 잠재 의미 공간으로 축소하는 대수적 정보검색 기법이다.
벡터공간 모델: 문헌과 질의를 용어 가중치 벡터로 표현하고 유사도로 관련성을 계산하는 대수적 모델이다.
확장 불리언 모델: 불리언 질의의 구조를 유지하면서 벡터공간 모델의 유사도 개념을 결합하여 부분일치와 순위화를 허용하는 모델이다.
확률 모델: 문헌이 질의에 적합할 확률을 추정하여 검색 결과를 순위화하는 모델이다.
퍼지집합 모델: 문헌이 어떤 개념 범주에 속하는 정도를 0과 1 사이의 부분 멤버십으로 나타내는 모델이다.

선지별 해설

① O — LSI는 용어-문헌 행렬에 특잇값 분해와 차원 축소를 적용하므로 대수적 모델에 속하는 개념적 기법이다.
② X — 틀린 표현은 '부분일치 검색은 허용하지 않는다'이다. 확장 불리언 모델은 전통적 불리언 모델의 엄격한 완전일치를 완화하여 부분일치를 허용하고 문헌을 관련성 순으로 배열할 수 있다.
③ O — 벡터공간 모델은 문헌과 질의를 벡터로 표현하는 대수적 모델이며, 벡터 간 유사도를 이용하므로 부분일치 검색과 순위화가 가능하다.
④ O — 확률 모델은 베이즈 정리를 바탕으로 적합·부적합 문헌에 대한 용어 출현 확률을 이용하여 문헌의 적합성 확률을 구할 수 있다.
⑤ O — 퍼지집합 모델은 소속 여부를 0 또는 1로만 나누지 않고, 한 문헌이 특정 색인어로 표현되는 개념 범주에 속하는 정도를 부분 멤버십으로 나타낸다.

NoteBase

오답 포인트

②는 전통적 불리언 모델의 완전일치 특성을 확장 불리언 모델에도 그대로 적용하게 만드는 함정이다. ‘확장’은 불리언 검색의 엄격성을 완화하여 부분일치와 순위화를 가능하게 한다는 뜻이다.

다시 나오면 이렇게 판단

모델마다 문헌을 무엇으로 표현하고 적합성을 어떻게 계산하는지 확인한다. 벡터·행렬이면 대수적 모델, 확률이면 확률 모델, 소속 정도이면 퍼지집합 모델이며, 확장 불리언 모델은 부분일치와 순위화를 허용한다.

암기 포인트

LSI는 행렬, 벡터공간은 유사도, 확률 모델은 적합 확률, 퍼지집합은 부분 멤버십, 확장 불리언은 부분일치·순위화이다.

NoteBase

14. 다음 <조건>에 따라 계산한 정보검색 평가척도의 값으로 옳은 것은?

< 조 건 >

ㄱ. 검색 대상 전체 문헌 수: 100개
ㄴ. 검색된 문헌 수: 20개
ㄷ. 검색된 문헌 중 검색자 요구에 적합한 문헌 수: 8개
ㄹ. 검색되지 않은 문헌 중 검색자 요구에 적합한 문헌 수: 60개

① 정확률은 20.0%이다.
② 재현율은 7.5%이다.
③ 잡음률은 80.0%이다.
④ 부적합률은 37.5%이다.
⑤ 누락률은 92.5%이다.

정답 4번

출제 포인트

이 문제는 검색 결과를 적합·부적합과 검색·미검색으로 나누고 정확률, 재현율, 잡음률, 부적합률, 누락률의 분모와 분자를 구별하는 문제이다.

조건 정리

전체 문헌은 100개이고 검색된 문헌은 20개이다. 검색된 적합 문헌은 8개이므로 검색된 부적합 문헌은 12개이다. 검색되지 않은 적합 문헌은 60개이므로 전체 적합 문헌은 68개이고, 전체 부적합 문헌은 32개이다. 따라서 검색되지 않은 부적합 문헌은 20개이다.

사용할 공식

정확률은 (검색된 적합 문헌)/(검색된 문헌)×100이다.
재현율은 (검색된 적합 문헌)/(전체 적합 문헌)×100이다.
잡음률은 (검색된 부적합 문헌)/(검색된 문헌)×100이다.
부적합률은 (검색된 부적합 문헌)/(전체 부적합 문헌)×100이다.
누락률은 (검색되지 않은 적합 문헌)/(전체 적합 문헌)×100이다.
각 척도는 검색 결과의 적합성 또는 전체 적합·부적합 문헌 중 검색된 비율을 측정하므로 해당 공식을 사용한다.

NoteBase

계산 과정

검색된 부적합 문헌 수는 $20 - 8 = 12$개이다. 전체 적합 문헌 수는 $8 + 60 = 68$개이고, 전체 부적합 문헌 수는 $100 - 68 = 32$개이다.
정확률은 $\frac{8}{20} \times 100 = 40.0\%$이다.
재현율은 $\frac{8}{68} \times 100 \approx 11.8\%$이다.
잡음률은 $\frac{12}{20} \times 100 = 60.0\%$이다.
부적합률은 $\frac{12}{32} \times 100 = 37.5\%$이다.
누락률은 $\frac{60}{68} \times 100 \approx 88.2\%$이다.

선지 판단

① X — 정확률은 검색된 문헌 20개 중 적합 문헌 8개의 비율이므로 40.0%이다. '20.0%'는 틀리고, 올바르게는 '정확률은 40.0%이다'이다.
② X — 재현율은 전체 적합 문헌 68개 중 검색된 적합 문헌 8개의 비율이므로 약 11.8%이다. '7.5%'는 틀리고, 올바르게는 '재현율은 약 11.8%이다'이다.
③ X — 검색된 부적합 문헌은 12개이므로 잡음률은 $\frac{12}{20} \times 100 = 60.0\%$이다. '80.0%'는 틀리고, 올바르게는 '잡음률은 60.0%이다'이다.
④ O — 전체 부적합 문헌은 32개이고 그중 검색된 부적합 문헌은 12개이므로 부적합률은 $\frac{12}{32} \times 100 = 37.5\%$이다.
⑤ X — 전체 적합 문헌 68개 중 검색되지 않은 적합 문헌은 60개이므로 누락률은 약 88.2%이다. '92.5%'는 틀리고, 올바르게는 '누락률은 약 88.2%이다'이다.

오답 포인트

핵심 함정은 분모의 선택이다. 정확률과 잡음률의 분모는 '검색된 문헌'이고, 재현율과 누락률의 분모는 '전체 적합 문헌'이며, 부적합률의 분모는 '전체 부적합 문헌'이다. 특히 부적합률을 잡음률과 혼동하면 12를 20으로 나누어 60.0%로 잘못 계산하게 된다.

다시 나오면 이렇게 판단

먼저 전체 문헌을 검색 여부와 적합 여부의 2 × 2 구조로 나누어 네 칸의 문헌 수를 완성한다. 그다음 척도 이름에 따라 분모 집단을 정하고, 해당 집단 가운데 목표 문헌의 비율을 계산한다.

암기 포인트

정확률·잡음률은 검색된 문헌이 분모이고, 재현율·누락률은 전체 적합 문헌이 분모이며, 부적합률은 전체 부적합 문헌이 분모이다. 정확률과 잡음률의 합, 재현율과 누락률의 합은 각각 100%이다.

NoteBase

15. n-그램(n-gram) 색인에 대한 설명으로 옳지 않은 것은?

① 형태소 분석이나 스테밍과 같은 언어학적 처리가 필요 없다.
② 띄어쓰기를 사용하지 않는 아시아권 문자를 색인하는 데 유용하다.
③ 질의어를 n-그램으로 변환하여 용어절단 검색을 할 수 있다.
④ 질의어와 색인어에 철자 오류가 있더라도 유사도가 높은 색인어를 검색할 수 있다.
⑤ 특정 언어의 문법에 독립적이기 때문에 색인의 크기가 줄어든다.

정답 5번

출제 포인트

n-그램 색인의 언어 독립성, 부분 문자열 검색, 철자 오류 대응이라는 장점과 색인 크기 증가라는 단점을 구분하는 문제이다.

핵심 개념

n-그램 색인은 문자열을 연속된 n개의 문자 단위로 겹쳐 나누어 각 조각을 색인하는 방식이다. 형태소나 단어 경계에 덜 의존하므로 다양한 언어에 적용하기 쉽고 부분 일치 검색에도 유리하지만, 한 문자열에서 많은 조각이 생성되므로 색인 항목과 저장 공간은 대체로 늘어난다.

선지별 해설

① O — 문자 단위 n-그램은 형태소 분석이나 어간 추출에 의존하지 않고 구성할 수 있으므로 별도의 언어학적 처리가 필수적이지 않다.
② O — 띄어쓰기나 명확한 단어 경계가 없는 문자 체계도 일정한 문자 길이로 분할할 수 있으므로 아시아권 문자 색인에 유용하다.
③ O — 질의어도 같은 길이의 n-그램으로 분해하여 색인과 대조하면 문자열의 일부가 일치하는 용어절단 검색을 지원할 수 있다.
④ O — 철자 오류가 있어도 질의어와 색인어가 여러 n-그램을 공유할 수 있으므로 공유 조각의 정도를 이용해 유사한 색인어를 검색할 수 있다.
⑤ X — 틀린 표현은 '특정 언어의 문법에 독립적이기 때문에 색인의 크기가 줄어든다'이다. 언어 독립성은 적용 범위의 장점일 뿐 색인 축소의 근거가 아니며, 올바르게는 하나의 문자열에서 서로 겹치는 여러 n-그램이 생성되므로 일반적으로 색인의 크기가 커질 수 있다고 해야 한다.

오답 포인트

⑤는 '언어 독립적'이라는 장점을 '색인 크기 감소'와 잘못 연결한다. n-그램은 형태소 분석을 줄일 수 있지만, 저장해야 할 중첩 조각의 수까지 줄이는 방식은 아니다.

NoteBase

다시 나오면 이렇게 판단

n-그램 선지는 '문자 단위 분할'에서 결과를 추론한다. 언어·단어 경계에 덜 의존하고 부분 일치·오류 허용에는 유리하지만, 중첩된 조각이 많이 생겨 색인 크기와 처리량은 증가할 수 있다고 판단한다.

암기 포인트

n-그램은 언어에는 독립적이고 오탈자에는 강하지만, 색인은 작아지는 것이 아니라 커질 수 있다.

NoteBase

16. with 연산자를 사용하는 검색시스템에서 색인어 A와 B가 다음과 같이 문헌에 출현하는 경우, 검색식 'A with B'로 검색되는 문헌은?

색인어	문헌번호	출현위치
A	1	7
	2	5
	3	2
	4	4
B	1	6
	2	6
	4	10
	5	19

① 1번 문헌
② 2번 문헌
③ 3번 문헌
④ 4번 문헌
⑤ 5번 문헌

정답 2번

출제 포인트

WITH 연산자는 두 색인어가 검색식에 적힌 순서대로 바로 이웃하여 출현하는 문헌을 찾는 인접 연산자이다. 표의 문헌번호별 출현위치를 대조하여 A 다음 위치에 B가 나오는 문헌을 찾는 문제이다.

자료 읽는 순서

1. A와 B가 함께 출현하는 문헌번호를 찾는다.
2. 같은 문헌에서 B의 출현위치가 A의 출현위치보다 정확히 1 큰지 확인한다.
3. 조건을 충족하는 문헌번호를 선지와 연결한다.

핵심 근거

2번 문헌에서 A의 출현위치는 5이고 B의 출현위치는 6이다. 따라서 A 바로 다음에 B가 출현하므로 검색식 'A with B'의 조건을 충족한다.

NoteBase

선지별 해설

① X — 1번 문헌에서는 B가 위치 6, A가 위치 7에 있어 순서가 ‘B A’이다. ‘A B’의 순서가 아니므로 검색되지 않는다. 올바르게는 1번 문헌이 아니라 A가 위치 5, B가 위치 6에 있는 2번 문헌이 검색된다.
② O — 2번 문헌에서는 A가 위치 5, B가 위치 6에 있어 두 색인어가 ‘A B’의 순서로 바로 이웃한다. 따라서 ‘A with B’로 검색된다.
③ X — 3번 문헌에는 A가 위치 2에 있지만 B의 출현 기록이 없다. 두 색인어가 함께 출현하지 않으므로 검색되지 않는다. 올바르게는 A와 B가 순서대로 인접한 2번 문헌이 검색된다.
④ X — 4번 문헌에는 A와 B가 모두 출현하지만 위치가 각각 4와 10이어서 바로 이웃하지 않는다. 올바르게는 단순히 같은 문헌에 있는 것만으로는 부족하고, A 다음 위치에 B가 있어야 한다.
⑤ X — 5번 문헌에는 B가 위치 19에 있지만 A의 출현 기록이 없다. 두 색인어가 함께 출현하지 않으므로 검색되지 않는다. 올바르게는 A와 B가 모두 있고 A 다음에 B가 인접한 2번 문헌이 검색된다.

오답 포인트

1번 문헌은 두 색인어가 서로 인접하므로 고르기 쉽지만 출현 순서가 ‘B A’로 반대이다. 또한 4번 문헌은 두 색인어가 함께 출현할 뿐 위치 4와 10으로 떨어져 있어 WITH의 순서·인접 조건을 충족하지 않는다.

다시 나오면 이렇게 판단

검색식이 ‘A with B’이면 각 문헌에서 A와 B의 출현 여부를 먼저 확인한 뒤, B의 위치가 A의 위치보다 정확히 1 큰 문헌만 남긴다.

암기 포인트

A with B는 ‘A 다음에 바로 B’이다.

NoteBase

17. 데이터 관리와 서비스를 위해 데이터 사서가 관심 가져야 할 요소로 옳지 않은 것은?

① 기관형 데이터 리포지터리
② 데이터 아카이빙 정책개발
③ 데이터와 문헌 연계 서비스
④ DMP(Data Management Plan) 작성지원 및 컨설팅
⑤ 데이터 이산화(discretization) 컨설팅

정답 5번

출제 포인트

데이터 사서가 담당하는 연구데이터의 수집·보존·공유·관리 지원과 데이터 분석 전처리 업무를 구별하는 문제이다.

핵심 개념

데이터 사서는 연구데이터의 체계적 관리와 장기 보존, 검색·접근·공유를 지원하는 정보전문가이다. 주요 업무에는 기관형 데이터 리포지터리 운영, 데이터 아카이빙 정책 개발, 데이터와 관련 문헌의 연계, DMP 작성 지원과 컨설팅이 포함된다. 반면 데이터 이산화는 연속형 수치를 일정한 구간이나 범주로 바꾸는 데이터 분석·전처리 기법이다.

선지별 해설

① O — 기관형 데이터 리포지터리는 기관이 생산한 연구데이터를 수집·보존하고 검색·공유하게 하는 기반이므로 데이터 사서의 관심 영역이다.
② O — 데이터의 선별, 보존 기간, 접근 권한과 폐기 원칙 등을 정하는 데이터 아카이빙 정책 개발은 연구데이터 관리 업무에 해당한다.
③ O — 데이터와 그 데이터를 생산하거나 활용한 논문 등 문헌을 연결하면 데이터의 맥락과 재이용 가능성이 높아지므로 데이터 사서가 제공할 수 있는 서비스이다.
④ O — DMP는 연구 과정에서 데이터의 생성·정리·보존·공유 방법을 미리 정하는 계획이므로, 그 작성 지원과 컨설팅은 데이터 사서의 대표적인 연구데이터 관리 서비스이다.
⑤ X — 틀린 표현은 '데이터 이산화 컨설팅'이다. 데이터 이산화는 연속형 값을 구간이나 범주로 변환하는 분석·전처리 기법으로, 데이터 사서의 일반적인 연구데이터 관리 서비스가 아니다. 올바르게는 데이터 관리·보존·공유 계획, 메타데이터 작성, 리포지터리 이용 등에 관한 컨설팅이 데이터 사서의 관심 영역이다.

NoteBase

오답 포인트

⑤에 '데이터'와 '컨설팅'이라는 말이 들어 있어 데이터 사서의 업무처럼 보이는 것이 함정이다. 그러나 이산화는 데이터를 관리·보존하는 활동이 아니라 값을 범주화하는 분석 전처리 기법이다.

다시 나오면 이렇게 판단

선지가 연구데이터의 수집·기술·보존·접근·공유 또는 관리계획 지원에 관한 것이면 데이터 사서의 업무로 판단한다. 통계 분석, 모델링, 값의 변환처럼 데이터 자체를 분석하는 기법이면 일반적인 데이터 사서 업무와 구별한다.

암기 포인트

데이터 사서는 데이터를 관리·보존·연결하고, 이산화는 데이터를 분석하기 위해 범주로 바꾼다.

NoteBase

18. 베이츠(M. Bates)는 정보학의 핵심 연구영역에서 '세 가지 큰 질문(Three Big Questions)'을 던지고 있다고 설명하였다. 다음 중 '세 가지 큰 질문'의 유형과 설명이 옳게 연결된 것은?

① 물리적 질문 - 기록된 정보세계의 특징과 법칙은 무엇인가?
② 설계 질문 - 사람들은 어떻게 정보를 찾고, 이해하고, 이용하는가?
③ 사회적 질문 - 정보의 사회적 효용성과 법적 가치는 무엇인가?
④ 물리적 질문 - 기록된 정보에 대한 가장 빠르고 효과적인 접근은 어떻게 이루어질 수 있는가?
⑤ 사회적 질문 - 정보사회를 어떻게 기술하고 설명하는가?

정답 1번

출제 포인트

이 문제는 베이츠(M. Bates)가 제시한 정보학의 세 가지 큰 질문인 물리적 질문, 사회적 질문, 설계 질문을 각각의 핵심 내용과 연결하는 문제이다.

핵심 정리

물리적 질문: 기록된 정보세계는 어떤 특징과 법칙을 가지는지를 탐구한다.
사회적 질문: 사람들은 정보와 어떻게 관계를 맺고, 정보를 어떻게 찾고 이해하며 이용하는지를 탐구한다.
설계 질문: 기록된 정보에 가장 빠르고 효과적으로 접근하게 하는 방법을 탐구한다.

선지별 해설

① O — '기록된 정보세계의 특징과 법칙은 무엇인가?'는 기록정보 세계 자체의 성질과 규칙을 다루므로 물리적 질문에 해당한다.
② X — 틀린 부분은 '설계 질문'이다. 사람들의 정보 탐색·이해·이용을 묻는 내용은 사회적 질문이다.
올바르게는 '사회적 질문 - 사람들은 어떻게 정보를 찾고, 이해하고, 이용하는가?'이다.
③ X — '정보의 사회적 효용성과 법적 가치는 무엇인가?'는 베이츠가 구분한 사회적 질문의 핵심 설명이 아니다.
올바르게는 '사회적 질문 - 사람들은 정보와 어떻게 관계를 맺고, 정보를 어떻게 찾고 이용하는가?'이다.
④ X — 틀린 부분은 '물리적 질문'이다. 기록된 정보에 가장 빠르고 효과적으로 접근하는 방법은 정보시스템과 서비스의 구축을 다루는 설계 질문이다. 올바르게는 '설계 질문 - 기록된 정보에 대한 가장 빠르고 효과적인 접근은 어떻게 이루어질 수 있는가?'이다.
⑤ X — '정보사회를 어떻게 기술하고 설명하는가?'는 베이츠의 사회적 질문을 정확히 나타내지 않는다.
올바르게는 '사회적 질문 - 사람들은 정보와 어떻게 관계를 맺고, 정보를 어떻게 찾고 이해하며 이용하는가?'이다.

NoteBase

출제자가 노린 함정

②의 '사람들이 정보를 찾고 이해하고 이용하는 방식'은 정보 이용을 더 잘 설계한다는 뜻으로 오인하기 쉽다. 그러나 사람의 정보행태 자체를 묻는 것은 사회적 질문이고, 접근 방법이나 시스템을 어떻게 만들 것인지를 묻는 것은 설계 질문이다.

다시 나오면 이렇게 판단

질문의 대상이 정보세계 자체의 법칙이면 물리적 질문, 사람의 정보행태이면 사회적 질문, 더 빠르고 효과적인 정보 접근 방법의 구현이면 설계 질문으로 판단한다.

암기 포인트

정보세계의 법칙 = 물리적, 사람의 정보행태 = 사회적, 빠르고 효과적인 접근 = 설계이다.

NoteBase

19. 클러스터링에 대한 설명으로 옳지 않은 것은?

① 비계층적 클러스터링 알고리즘은 미리 생성할 클러스터의 기준치를 결정하므로 경험적 기법이라고도 한다.

② 단일연결 기법은 계층적 클러스터링 알고리즘으로 통합 대상이 되는 두 클러스터에서 가장 유사한 두 개 구성원의 유사도를 기반으로 클러스터를 생성하는 기법이다.

③ 계층적 클러스터링 알고리즘은 처리 시간 측면에서 비계층적 기법에 비하여 효율성이 떨어진다.

④ 완전연결 기법은 계층적 클러스터링 알고리즘으로 통합 대상이 되는 두 클러스터에서 가장 유사하지 않은 두 개 구성원의 유사도를 기반으로 클러스터를 생성하는 기법이다.

⑤ K-means 클러스터링 기법은 계층적 클러스터링 알고리즘으로 임의로 형성된 K개의 초기 클러스터 집합에서 시작된다.

정답 5번

출제 포인트

비계층적 클러스터링과 계층적 클러스터링의 차이, 그리고 단일연결법·완전연결법·K-means의 분류와 결합 기준을 묻는 문제이다.

핵심 개념

비계층적 클러스터링은 군집 수를 미리 정하고 개체를 군집에 배정·재배정하는 방법이며, K-means가 대표적이다.
계층적 클러스터링은 가까운 군집을 차례로 합치거나 큰 군집을 나누어 계층 구조를 만드는 방법이다.
단일연결법은 두 군집에서 가장 가까운 구성원 쌍을, 완전연결법은 가장 먼 구성원 쌍을 군집 간 거리의 기준으로 삼는다.

NoteBase

선지별 해설

① O — 비계층적 클러스터링은 생성할 클러스터 수를 사전에 정해야 하므로 분석자의 경험적 판단이 개입될 수 있다.
② O — 단일연결법은 두 군집 사이에서 가장 유사한 구성원 쌍, 즉 거리가 가장 가까운 쌍을 기준으로 군집을 결합한다.
③ O — 계층적 클러스터링은 군집 간 거리를 반복적으로 계산하고 결합 관계를 갱신하므로 일반적으로 비계층적 기법보다 계산 부담이 크고 처리 시간의 효율성이 낮다.
④ O — 완전연결법은 두 군집 사이에서 가장 유사하지 않은 구성원 쌍, 즉 거리가 가장 먼 쌍을 기준으로 군집 간 거리를 정한다.
⑤ X — 틀린 표현은 'K-means 클러스터링 기법은 계층적 클러스터링 알고리즘'이다. K-means는 K개의 초기 중심을 정한 뒤 각 자료를 가까운 중심에 배정하고 중심을 갱신하는 과정을 반복하는 비계층적 분할 클러스터링 기법이다.

오답 포인트

⑤에는 'K개의 초기 클러스터에서 시작한다'는 맞는 특징이 섞여 있지만, 이를 근거로 K-means를 계층적 기법으로 분류하면 안 된다. 초기 군집 수 K를 정해 반복 재배정하는 방식은 비계층적 기법의 핵심 특징이다.

다시 나오면 이렇게 판단

군집 수 K를 먼저 정하고 소속과 중심을 반복 갱신하면 비계층적 기법으로 판단한다. 군집을 단계적으로 합치거나 나누어 계층 구조를 만들면 계층적 기법으로 판단한다. 연결법은 단일연결이 가장 가까운 쌍, 완전연결이 가장 먼 쌍을 기준으로 삼는다고 구분한다.

암기 포인트

K-means는 비계층적 분할법, 단일연결은 최단거리, 완전연결은 최장거리이다.

NoteBase

20. 정보검색 시스템의 성능 평가척도들 중 특정 질의에 대한 전체 적합 문헌 수를 알지 못하더라도 사용 가능한 척도만을 <보기>에서 모두 고르면?

< 보 기 >

ㄱ. 정확률	ㄴ. 재현율
ㄷ. 잡음률	ㄹ. 누락률
ㅁ. E 척도	ㅂ. F 척도
ㅅ. R-정확률(R-precision)	ㅇ. 순위정확률(precision at k)

① ㄱ, ㄷ, ㅁ
② ㄱ, ㄷ, ㅅ
③ ㄱ, ㄷ, ㅇ
④ ㄴ, ㄹ, ㅂ
⑤ ㄴ, ㄹ, ㅅ, ㅇ

정답 3번

출제 포인트

검색 성능척도의 분모에 전체 적합 문헌 수가 들어가는지를 구별하는 문제이다.

핵심 개념

정확률은 검색된 문헌 중 적합 문헌의 비율이고, 잡음률은 검색된 문헌 중 부적합 문헌의 비율이므로 전체 적합 문헌 수가 없어도 계산할 수 있다. 순위정확률은 상위 k개 검색 결과 중 적합 문헌의 비율이므로 역시 전체 적합 문헌 수를 요구하지 않는다. 반면 재현율과 누락률은 전체 적합 문헌을 기준으로 하며, E 척도와 F 척도는 정확률과 재현율을 함께 반영한다. R-정확률은 전체 적합 문헌 수 R만큼 검색했을 때의 정확률이므로 R을 알아야 한다.

NoteBase

선지별 해설

① X — ㄱ 정확률과 ㄷ 잡음률은 사용할 수 있지만, ㅁ E 척도는 재현율을 반영하므로 전체 적합 문헌 수가 필요하다. 틀린 부분은 ㅁ을 포함한 것이며, 올바른 조합은 ㄱ, ㄷ, ㅇ이다.
② X — ㄱ 정확률과 ㄷ 잡음률은 사용할 수 있지만, ㅅ R-정확률은 전체 적합 문헌 수 R을 알아야 상위 R개를 정할 수 있다. 틀린 부분은 ㅅ을 포함한 것이며, 올바른 조합은 ㄱ, ㄷ, ㅇ이다.
③ O — ㄱ 정확률, ㄷ 잡음률, ㅇ 순위정확률은 각각 검색된 문헌 수 또는 미리 정한 k를 분모로 하므로 전체 적합 문헌 수를 몰라도 계산할 수 있다.
④ X — ㄴ 재현율과 ㄹ 누락률은 전체 적합 문헌 수를 기준으로 하고, ㅂ F 척도도 재현율을 반영하므로 모두 전체 적합 문헌 수가 필요하다. 틀린 부분은 ㄴ, ㄹ, ㅂ을 사용 가능하다고 묶은 것이며, 올바른 조합은 ㄱ, ㄷ, ㅇ이다.
⑤ X — ㅇ 순위정확률만 전체 적합 문헌 수 없이 사용할 수 있다. ㄴ 재현율과 ㄹ 누락률은 전체 적합 문헌 수를 분모로 하고, ㅅ R-정확률은 R 자체가 전체 적합 문헌 수이므로 사용할 수 없다. 틀린 부분은 ㄴ, ㄹ, ㅅ을 포함한 것이며, 올바른 조합은 ㄱ, ㄷ, ㅇ이다.

오답 포인트

정확률과 재현율의 분모를 뒤바꾸거나, R-정확률의 R을 임의의 검색 개수로 오해하기 쉽다. 이 문항에서는 척도가 전체 적합 문헌 수 또는 그에 기반한 재현율을 사용하는지만 확인해야 한다.

다시 나오면 이렇게 판단

분모가 검색된 문헌 수나 미리 정한 k이면 전체 적합 문헌 수 없이 계산할 수 있다. 분모가 전체 적합 문헌 수이거나 척도에 재현율 또는 R이 들어가면 전체 적합 문헌 수가 필요하다고 판단한다.

암기 포인트

전체 적합 문헌 수 없이 가능한 것은 정확률·잡음률·순위정확률이고, 재현율·누락률·E·F·R-정확률은 불가능하다.

NoteBase